Interdependent Model for Point-of-Interest Recommendation via Social Networks

Author

Hashim-Jones, Jake, Wang, Can, Islam, Md Saiful, Stantic, Bela

Published

2018

Journal Title

Lecture Notes in Computer Science

Version

Accepted Manuscript (AM)

DOI

10.1007/978-3-319-92013-9_13

Rights statement

Interdependent Model for Point-of-Interest Recommendation Via Social Networks

Jake Hashim-Jones, Can Wang, Md. Saiful Islam, Bela Stantic

School of Information and Communication Technology, Griffith University, Australia
jake.hashim-jones@griffithuni.edu.au
{can.wang, saiful.islam, b.stantic}@griffith.edu.au

Abstract. Point-of-Interest (POI) recommendation is an important way to help people discover attractive places. POI recommendation approaches are usually based on collaborative filtering methods, whose performances are largely limited by the extreme scarcity of POI check-ins and a lack of rich contexts, and also by assuming the independence of locations. Recent strategies have been proposed to capture the relationship between locations based on statistical analysis, thereby estimating the similarity between locations purely based on the visiting frequencies of multiple users. However, implicit interactions with other link locations are overlooked, which leads to the discovery of incomplete information. This paper proposes a interdependent item-based model for POI recommender systems, which considers both the intra-similarity (i.e. co-occurrence of locations) and inter-similarity (i.e. dependency of locations via links) between locations, based on the TF-IDF conversion of check-in times. Geographic information, such as the longitude and latitude of locations, are incorporated into the interdependent model. Substantial experiments on three social network data sets verify the POI recommendation built with our proposed interdependent model achieves a significant performance improvement compared to the state-of-the-art techniques.

1 Introduction

Location-based social networks have become increasingly popular in the past few years. Popular networks such as Foursquare, WhatsApp, FaceBook, Snapchat and even Instagram all include check-in functionality, allowing users to share their location-based information with their friends. In 2016, the social network Foursquare had over 50 million users and 8 billion POI check-ins [1]. A point of interest (POI) is essentially a locational hotspot, such as a restaurant, shopping center or public park. With the emergence of location-based social networks, POI recommender systems have been developed to recommend locations for visit [1].

Compared with traditional recommendation problems (e.g. product and movie), there are a number of key performance-limiting challenges that current POI recommender systems face. One of the prominent issues is the scarcity of available data. Recommender systems generally work by analyzing a user's activities with respect to items (which in this case is POIs) [2]. A key issue with POI data in particular, is the high scarcity. A set of POIs can be extensive, however users only tend to visit a limited number of them [1]. As a result, there is a limited amount of POI visitation data available that a recommender system can use to model a user's preferences and behavior. Another challenging issue with POI recommender systems is that generally, they are incapable of dealing with rich contexts (e.g. geographical and social data), and base their predictions solely on the visiting frequencies [2]. As visiting frequency tends to be scarce, it is important that the influence of additional rich contexts on user behavior is considered.

Three popular methods used to power recommender systems are: item-based and user-based collaborative filtering, and matrix factorization techniques. The item-based models consider the similarity between locations based on common users that have visited them [3]. User-based models involve generating user predictions based on similarities between users who have visited similar locations [3]. Matrix factorization works by decomposing user visitation data into various latent factors of users and locations [4]. These methods tend to be subject to the same issues that all POI recommender systems face: data scarcity and a lack of rich contextual awareness. Another factor that is not usually considered is the coupled relationship between locations. Some research work [1] has analyzed the explicit statistical relationship for locations based on check-ins. However, they do not address the underlying implicit relationship, thus failing to capture the complete semantic relationship between locations.

This study aims to address the challenge of capturing the coupled relationship between locations, and implementing them in a POI recommender system. The purpose is to address the data scarcity problem by introducing more information, as well as taking more rich context into account by geographical data. In this work, we propose a novel approach to measure the relationship between locations by capturing both the intra-similarity and inter-similarity, using the co-occurrence and link between them, respectively. The interdependent model is based on statistical analysis of location occurrence patterns to determine how similar locations are (intra-similarity), based on check-ins. We also attempt to capture the relationship between locations that are not visited by the same users (inter-similarity), via implicit links. Throughout this paper, we use interdependent and coupled interchangeably. The key contributions are as follows:

- **Interdependent Similarity of Locations**: We capture the coupled relationship between multiple locations with respect to both intra-similarity and inter-similarity, via modelling the co-occurrences and links of locations.
- **Geographical Information**: We also integrate explicit rich context data in the form of geographical distance using longitude and latitude, whilst capturing the implicit coupled relationship between locations. This results in a successful implementation of the interdependent item-based model.
- **Excellent Performance on Real-Life Social Networks**: We run our model on the real-life data, including the world-wide Gowalla data, Australia Gowalla data, and a closer case-study in Gold Coast via Twitter. We observe major performance increases in terms of accuracy and error reduction. In some instances, the increase exceeds 32-fold (3118%). It verifies the effectiveness of applying the coupled similarity and geographic information.

2 Related Work

Content-based filtering (CBF) and collaborative filtering (CF) techniques are two popular approaches for recommender systems. CBF [2] approaches produce recommendations for users by analyzing the content of items and user profiles. CF [2] methods produce recommendations or predictions relying solely on the past activities of users. CF can be classified into two classes: memory-based and model-based algorithms. Memory-based algorithms are further divided into user-based and item-based methods under different hypothesis [3]. User-based approaches assume that similar users have similar tastes and make recommendations based on the opinion of users' most similar neighbors [3]. On the other hand, item-based methods assume that users are interested in similar items [3]. Unlike memory-based approaches, which use the entire user-item matrix to make recommendations, the model-based approaches [4] learn a predictive model from training data to predict. Some of the most successful realizations of model-based algorithms are built upon matrix factorization [3], due to their good scalability and predictive accuracy. Matrix factorization techniques assume that users' preferences and items' characteristics are described by latent factors [4].

A few more recent and advanced methods include the Poisson factor models [1], deep learning models [5], and probabilistic models [6], and etc. However, none of them consider the coupling relationship between locations in the POI recommendation. In our proposed interdependent model, the basic item-based method is used to incorporate the coupling relationship to determine if it produces an improved prediction. We may consider to apply the interdependent model to integrate with the sophisticated approaches as the future work, since our proposed interdependent similarity is capable to be widely fit to different model settings as required.

3 Problem Statement

A large number of users with multiple locations can be represented by the user-location frequency graph, which is a weighted undirected bipartite graph $G^f_{ul} = (U, L, E_{ul}, W^f_{ul})$, where U is a set of vertices that represent users, L is a collection of nodes that represent locations, E_{ul} is a set of edges to denote visits, and the edge weights W^f_{ul} specify the number of visits to a location by a user. Given m users $U = \{u_1, \cdots, u_m\}$ and n locations $L = \{l_1, \cdots, l_n\}$, we then can define an adjacency matrix $M^f_{ul} = [x_{si}] \in \mathbb{R}^{m \times n}$ for the frequency graph G^f_{ul}. Each entry x_{si} in M^f_{ul} quantifies how many times the user u_s visits each location l_i. For example, Fig. 1a represents a tabular form of an adjacency matrix, referred to as a "user-location frequency matrix", in which m users are each represented by rows, and n locations are denoted by columns. The entry for a user u_s and a location l_i reflects the frequency that the user has visited the location, i.e. f_{u_s,l_i}.

The POI recommendation problem is to predict the top-K ranking locations for the user u_s with the estimated check-in entries (prediction scores) p_{u_s,l_i}, which is predicted via analyzing the historical visiting locations. For example, Fig. 1b shows a tabularized matrix which now contains prediction scores (p_{u_s,l_i}) for each user (u_s) for visiting each location (l_i) (the ideal output of a system addressing the POI recommendation problem). In this example, u_3 shows a series of numbers (prediction scores) for each location. A higher prediction score indicates that the system predicts the user more likely to visit that location. Hence, in this example, if $K = 2$, the top-2 locations would be l_3 and l_5.

		Locations				
		l_1	l_2	l_3	...	l_n
Users	u_1	f_{u_1,l_1}	f_{u_1,l_2}	f_{u_1,l_3}	...	f_{u_1,l_n}
	u_2	f_{u_2,l_1}	f_{u_2,l_2}	f_{u_2,l_3}	...	f_{u_2,l_n}
	u_3	f_{u_3,l_1}	f_{u_3,l_2}	f_{u_3,l_3}	...	f_{u_3,l_n}
	...	...	...	...	...	...
	u_m	f_{u_n,l_1}	f_{u_n,l_2}	f_{u_m,l_3}	...	f_{u_m,l_n}

(a) User-Location Frequency Matrix

		Locations				
		l_1	l_2	l_3	l_4	l_5
Users	u_1	p_{u_1,l_1}	p_{u_1,l_2}	p_{u_1,l_3}	p_{u_1,l_4}	p_{u_1,l_5}
	u_2	p_{u_2,l_1}	p_{u_2,l_2}	p_{u_2,l_3}	p_{u_2,l_4}	p_{u_2,l_5}
	u_3	**0.3**	**1.2**	**4.7**	**1.8**	**3.6**
	...	...	...	...	...	...
	u_m	p_{u_n,l_1}	p_{u_n,l_2}	p_{u_m,l_3}	p_{u_m,l_4}	p_{u_m,l_5}

(b) User-Location Prediction Matrix

Fig. 1. Example user-location matrices represented in tabular form.

4 Interdependent Item-Based Model

4.1 Baseline Model

The basic model of the recommender system used is a classic example of item-based collaborative filtering [3], in which the items are POIs in the POI recommendation settings. In this model, a similarity matrix is calculated, representing the similarity between different POI locations. This matrix is based on cosine similarity between different location pairs [3]. The cosine similarity can then be used to calculate predicted frequency values based on Equation (1) [7]. In this equation, the predicted frequency ($f_{s,i}$) of a user (u_s) for a location (l_i) is equivalent to the ratio of the sum of the similarities between l_i and all other locations multiplied by the frequency of visits for u_s at each location, and the sum of the absolute similarities between l_i and all other locations, as outlined below:

$$\hat{f}_{s,i} = \frac{\sum_{l_j} S_l^{cos}(l_i, l_j)(f_{s,j})}{\sum_{l_j} |S_l^{cos}(l_i, l_j)|}, \tag{1}$$

where $\hat{f}_{s,i}$ is the prediction score for user u_s at location l_i, and $S_l^{cos}(l_i, l_j)$ is the cosine similarity between l_i and l_j. Equation (1) is used to yield a matrix of predictions between each user and each location. Upon running data through the above processes, it is possible to obtain a prediction score for each user-location pair. It can then be added to a prediction matrix, as in Fig. 1(b). The baseline model only considers the explicit relationship between locations via check-ins.

4.2 Interdependent Model

The interdependent model addresses the coupling relationship among locations based on the user-location TF-IDF matrix, as well as geographic information.

User-Location TF-IDF Matrix The user-location TF-IDF matrix (M_{ul}^{ti}) is derived from the user-location frequency matrix, by following an idea from text mining [8], where each document is regarded as a user, each term is regarded as a location, and the frequency of the word in an document is regarded as the frequency of location visits by a user. So in such a matrix, each row represents a user, while each column represents a location. The elements of the matrix are the TF-IDF values of each user-location pairing. As a result, the values along the matrix diagonal will always be 0. The TF in M_{ul}^{ti} can be calculated as the ratio of the location visit frequency ($f_{l_i u_s}$) of a user and the total number of locations visited by the user ($\sum_{l_i' \in u_s} f_{l_i,u_s}$). The IDF in M_{ul}^{ti} can be calculated as the natural logarithm (for scaling purposes) of the ratio between the total number of users and the number of users that have visited a specific location ($|u \in U : l_i \in u|$). TF-IDF value is the product of TF and IDF, formalized as:

$$tf \cdot idf = \frac{f_{l_i,u_s}}{\sum_{l_i' \in u_s} f_{l_i,u_s}} \cdot \ln \frac{|U|}{|u \in U : l_i \in u|}. \tag{2}$$

The Location-Location Transition Matrix For the geographical information, we include a location-location transition matrix in our proposed model. The rows and columns represent a set of POIs. Thus, every element of the matrix represents the geographical distance between a pair of POIs (l_i and l_j) taken from the set. This data can be calculated as the numerical form of Euclidean distance from geographical co-ordinates, by using longitude and latitude. We use such geographic information to define the coupled relationship among locations.

Based on the user-location TF-IDF matrix and the geographic information in the form of location-location transition matrix, we are ready to propose the coupled similarity between locations via intra-similarity and inter-similarity.

Intra-Similarity of Locations The first step in measuring the semantic relatedness between locations is to explore the explicit semantic relationship between them, referred to as the intra-similarity of locations.

The intra-similarity between locations is based on the statistical analysis of location occurrence patterns [1]. It supposes that locations are relational if they are visited by the same user, and the intra-similarity between locations can be estimated by calculating their co-occurrence frequency across all users. In most of the previous approaches, the relationship between locations is simply estimated by the inner product of their distribution across the entire user set. Here, we adapt the popular co-occurrence measure Jaccard [9] to capture the intra-similarity between locations according to their association in the user sets. We capture the intra-relationship between locations for a single user, and then extended it to compute the intra-coupling across the whole user set.

The intra-similarity of locations matrix (δ^{Ia}) is derived from the user-location TF-IDF matrix. Both the rows and columns of this matrix represent locations of the data set and each element represents the intra-similarity of the respective location pair. At a user-level, intra-similarity is calculated for user u_s as follows.

$$\delta^{Ia}(l_i, l_j | u_s) = (x'_{si} \cdot x'_{sj})/(x'_{si} + x'_{sj}), \tag{3}$$

where x'_{si} and x'_{sj} represent TF-IDF values for a specific user at l_i and l_j, respectively. To move intra-similarity of locations to the corpus level, the average intra-similarity is taken for the set of users that have visited either location. However, only the users that have visited both locations contribute to this total calculation (other-wise their contribution to the sum is 0), as defined below.

$$\delta^{Ia}(l_i, l_j|U) = \frac{1}{|S|} \sum_{u_s \in S'} \delta^{Ia}(l_i, l_j|u_s), \tag{4}$$

where $S = \{u_s|(x'_{si} \neq 0) \vee (x'_{sj} \neq 0)\}$ (the set of users visiting either location) and $S^{'} = \{u_s|(x'_{si} \neq 0) \wedge (x'_{sj} \neq 0)\}$ (the set of users visiting both locations).

When constructing the final matrix, if l_i and l_j are the same, then their intra-similarity will always be 1, otherwise, the remaining intra-similarities are normalized via feature scaling. This process is shown in Equation (5).

$$\delta^{Ia}(l_i, l_j) = \begin{cases} 1 & \text{if } i = j, \\ \frac{\delta^{Ia}(l_i,l_j|U) - \min\limits_{1 \leq i,j \leq n} \{\delta^{Ia}(l_i,l_j|U)\}}{\max\limits_{1 \leq i,j \leq n} \{\delta^{Ia}(l_i,l_j|U)\} - \min\limits_{1 \leq i,j \leq n} \{\delta^{Ia}(l_i,l_j|U)\}} & \text{if } i \neq j, \end{cases} \tag{5}$$

where $\delta^{Ia}(l_i, l_j)$ is the intra-similarity between l_i and l_j, $\delta^{Ia}(l_i, l_j|U)$ represents the corpus-level intra-similarity between those locations before normalization.

Inter-Similarity of Locations The above intra-similarity only captures the relationship between locations based on their co-occurrence frequency. However, some locations are related to each other even though they do not happen to be visited by the same user. As the absence of this underlying relationship, the semantic information in the original locations cannot be completely captured, which brings difficulties into the recommender system. Hence, we propose to define the inter-similarity of locations to capture this underlying relationship.

We have captured the intra-similarity $\delta^{Ia}(l_i, l_k)$ between locations l_i and l_k, and $\delta^{Ia}(l_k, l_j)$ between locations l_k and l_j. The $\delta^{Ia}(l_i, l_k)$ indicates the possibility of visiting both locations l_i and l_k. Similarly, the $\delta^{Ia}(l_k, l_j)$ captures the possibility of visiting both locations l_k when l_j. Intuitively, it is easy to see that, $\delta^{Ia}(l_i, l_k)$ and $\delta^{Ia}(l_k, l_j)$ should imply the possibility to visit both locations l_i and l_j, thus the possibility can be determined using them.

The inter-similarity of locations matrix (δ^{Ie+geo}) compares the closeness of locations even when they have not been visited by the same users. Equation (6) is used to calculate the inter-similarity between locations. This matrix is derived based on the values of the intra-similarity of locations matrix. The process also involves taking geographic similarity into account, hence the information from the location-location transition matrix is used to build the weights for locations.

$$\delta^{Ie+geo}(l_i, l_j) = \max_{l_k \in L - \{l_i, l_j\}} \{\min\{g_{ik} \cdot \delta^{Ia}(l_i, l_k), g_{kj} \cdot \delta^{Ia}(l_k, l_j)\}\}, \tag{6}$$

where δ^{Ia} is the intra-similarity between locations defined in Equation (5). g_{ik} is the geographic similarity between locations l_i and l_k, and g_{ik} is obtained via transforming the normalized max-min distance $\hat{y}_{kj}$ to similarity, by applying

$g_{ik} = 1/(1 + \hat{y}_{ik})$. The value of $\hat{y}_{ik}$ is calculated as in Equation (7):

$$\hat{y}_{ik} = [y_{ik} - \min_{1 \le i,k \le n} \{y_{ik}\}]/[\max_{1 \le i,k \le n} \{y_{ik}\} - \min_{1 \le i,k \le n} \{y_{ik}\}], \tag{7}$$

where y_{ik} is the (i, k) entry in the location-location transition matrix based on Euclidean distance. The same calculation process applies to g_{kj} as well. For $i = j$, we set $\delta^{Ie+geo}(l_i, l_j) = 1$. Otherwise, $\delta^{Ie+geo}(l_i, l_j)$ is used.

Interdependent Similarity of Locations The intra-similarity discovers the explicit relationship between locations using their co-occurrences, and the inter-similarity discovers the implicit relationship by their interaction with other locations. Hence, to capture the complete semantic relationship between locations to build a high-quality recommender system, we couple the intra-similarity and inter-similarity together to measure the semantic relatedness between locations.

The final matrix is the interdependent location similarity matrix (δ^C), which is representative a graph derived from the inter-similarity of locations matrix and intra-similarity of locations matrix. This matrix is reflective of both the inter-similarity and intra-similarity between locations. During the coupling, the balance between intra-similarity and inter-similarity is controlled by the parameter α. α is between 0 and 1, with higher values giving more precedence to inter-similarity and lower values giving more precedence to intra-similarity. The equation used to calculate the interdependent similarity of locations is outlined as below.

$$\delta^C(l_i, l_j) = \begin{cases} 1 & \text{if } i = j, \\ (1 - \alpha) \cdot \delta^{Ia}(l_i, l_j) + \alpha \cdot \delta^{Ie}(l_i, l_j) & \text{if } i \neq j, \end{cases} \tag{8}$$

where $\delta^C(l_i, l_j)$ represents the interdependent location similarity between l_i and l_j. If l_i and l_j are the same, $\delta^C(l_i, l_j)$ is set to 1. The value of $\delta^C(l_i, l_j)$ falls within [0, 1], in which 0 indicates that two locations are completely unrelated, while 1 indicates that they have the highest relatedness. Hence, the higher the coupled score, the more similar the two locations. In our proposed interdependent model, we replace the cosine similarity from Equation (1) with coupled similarity, to calculate the recommendation score.

5 Empirical Study

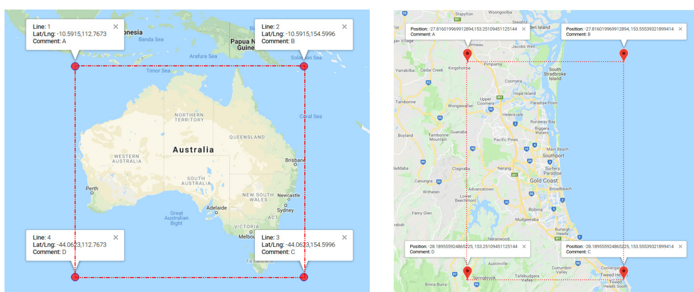
5.1 Data Sets

In this study, a total of three data sets were used. Two of these were derived from the public Gowalla data set. This data set included world-wide check-in data from February 2009 to October 2010. This data set was sourced from Liu et al. (2017) [1], who also filtered out users with fewer than 15 unique location check-ins and POIs with fewer than 10 visitors.

- **Gowalla**: The Gowalla data set used in this study is sometimes referred to as "Gowalla (0.1)". This is a version of the public Gowalla data set (mentioned above) that has been scaled down to only include the first 10% of users and first 10% of locations. This scaled version of the data set contains 1873 users, 3250 locations and 51330 check-ins. The data set has a sparsity of 99.16%.

- **Gowalla Australia**: This is referred to as the "Gowalla (Aus)" data set. It is filtered from the original public Gowalla data set obtained by [1], which again contains world-wide check-in data from February 2009 to October 2010. The filtered version of this data set contains check-ins to locations only within Australia. The data is also further filtered to remove users with less than 10 check-ins and locations with less than 10 visitors. The remaining data set consists of 7,717 check-ins from 866 users at 1,118 locations. The sparsity of this data set is 99.20%.
- **Gold Coast Twitter**: This data set (referred to as "GC Twitter") is obtained by cross-referencing a set of all tweets made in the Gold Coast with a set of known POIs obtained by crawling Foursquare. This is done by matching the geo-locations of tweets to the geo-locations of Foursquare POIs. The tweets were made between 1 January 2017 and 31 December 2017. The final data set is further filtered down to remove users with less than 5 tweet locations and locations with less than 5 users making tweets. The final GC twitter data set contains 31730 check-ins of 1529 different users at 786 Gold Coast POI locations. The sparsity of this data set is 99.30%.

The geographical bounding boxes of the Gowalla (Aus) and GC Twitter data sets are shown in Fig. 2. Training and testing data sets are obtained by randomly partitioning the frequencies in User-POI check-in matrix. 80% of the data is used to train the models and the remaining 20% is used to evaluate.



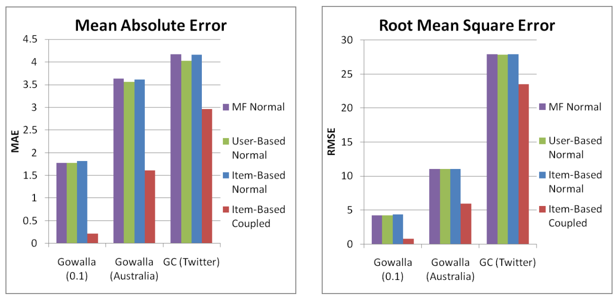
(a) Gowalla (Aus) Data Area (b) GC Twitter Data Area

Fig. 2. The geographical areas for (a) Gowalla (Aus) and (b) GC Twitter data sets.

5.2 Performance Metrics and Experimental Settings

Mean Error Two mean error metrics are used to evaluate the models. MAE and RMSE are both metrics that are used for evaluating absolute error [10]. For the existing methods, we use the user-location matrix as truth. For the interdependent model however, in order to address scaling issues, a TF-IDF matrix is applied as truth. The smaller the MAE and RMSE error scores, the better the recommendation quality.

Precision and Recall Precision (Pre@K) and recall (Rec@K) are used to further evaluate the models. This involves calculating precision and recall in the top-K most recommended POIs. In this evaluation, precision is the ratio of positive predictions and the number of predicted recommendations [1]. Recall is the ratio of positive predictions and the number of POIs actually visited [1]. Since precision and recall are expected to display different behaviors with different values of K, the F1-score [11] is also used as a means of coupling precision and recall into a single metric. The larger the precision, recall, and F1 scores, the better the recommendation quality.



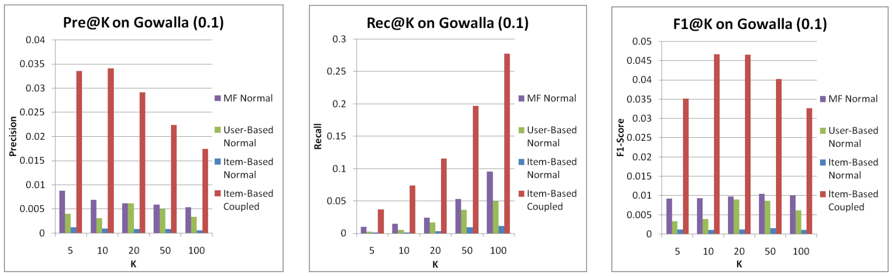
(a) MAE (b) RMSE

Fig. 3. Mean error in all data sets.

Experimental Settings For comparative purposes, all three data sets are also run through three basic models: item-based collaborative filtering [3], user-based collaborative filtering [3], and matrix factorization (MF) using singular value decomposition [4]. There are no relevant parameters for the basic models. For the interdependent model, the value of α is set to 0.5 (giving equal weighting to both intra-similarity and inter-similarity when determining interdependent location similarity). Each figure in the following section represents the proposed interdependent model (the "Item-Based Coupled" model) using the red bar. All experiments are run on the *Griffith University Gowonda HPC Cluster* using Intel Xeon CPU X5650 processors@2.67GHz.

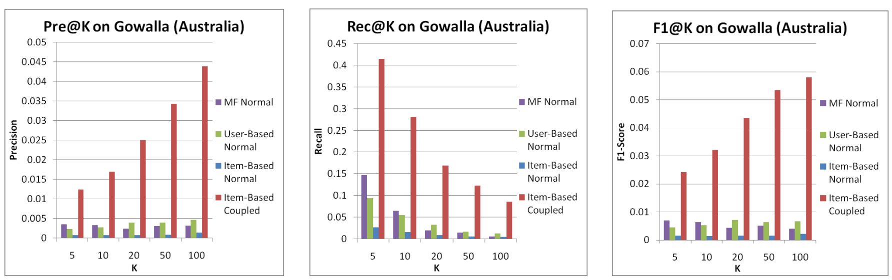
5.3 Experimental Results

With respect to mean error (in both metrics) all values between all of the basic models are similar (as shown in Fig. 3). However, there is a very clear and distinctive drop in MAE and RMSE with the interdependent model when comparing it with the basic models. The most prominent drop occurrs in the Gowalla (0.1) data set (which overall also has the lowest mean error for all models). The change (between the interdependent model and the basic model) in MAE is approximately -8.30-fold (an 87.95% decrease) and in RMSE, -5.58-fold (an 82.09% decrease). The second greatest drop in mean error occurs with the Gowalla (Aus) data set. For this data set, the change (between the interdependent item-based model and the basic item-based model) in MAE is approximately -2.24-fold (a 55.34% decrease) and the change in RMSE is approximately -1.86-fold (a 46.19% decrease). For the GC Twitter data set, the fold change (between the interdependent item-based model and the basic item-based model) in MAE is approximately -1.40 (a drop of 28.82%) whilst the change in RMSE is approximately -1.19 fold (a drop of 15.80%). This is the lowest drop in mean error. In terms of mean error, the interdependent model is thus considered a tremendous success as it reduces error significantly across all data sets.



(a) Pre@K Gowalla 0.1 (b) Rec@K Gowalla 0.1 (c) F1@K Gowalla 0.1

Fig. 4. Precision, recall and F1-scores on the Gowalla (0.1) data set.

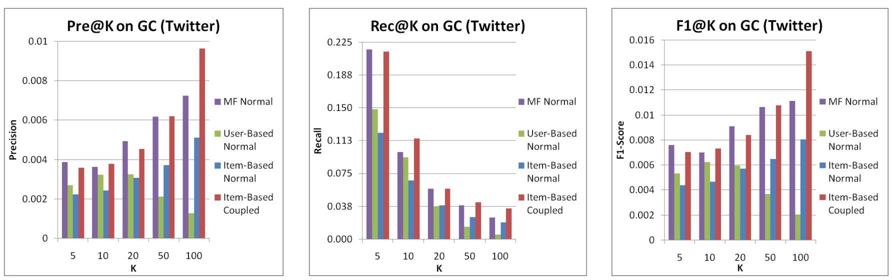


(a) Pre@K Gowalla Aus (b) Rec@K Gowalla Aus (c) F1@K Gowalla Aus

Fig. 5. Precision, recall and F1-scores on the Gowalla (Aus) data set.

The precision and recall data also shows a prominent increase in performance for both metrics with the interdependent model on the all data sets. There is a remarkable increase in precision, recall and F1-scores which usually supersedes all other models significantly. On all data sets, the basic item-based model tends to perform worse than all other models, however, the interdependent item-based model usually supersedes all other models. The most prominent changes in precision/recall metrics are observed when analyzing the Gowalla (0.1) data set (Fig. 4). Across all values of K, the average precision increase (between the interdependent item-based model and the basic item-based model) is approximately 31.68-fold (a 3068.68% increase). The average increase in recall is 31.06-fold (a 3005.86% increase) and similarly, the average increase in the F1-scores is 32.18-fold (a 3118.38% increase). The changes are also significant in the Gowalla (Aus) data set (Fig. 5). Again, across all values of K, the average precision increase (between the interdependent item-based model and the basic item-based model) is approximately 27.31-fold (a 2631.41% increase). Recall and F1-scores, respectively, experience approximate 19.90-fold and 25.74-fold increases. This is equivalent to increases of 1890.33% and 2473.89%.

The poorest precision and recall scores are observed with the GC Twitter data set (Fig. 6), across all values of K, the average precision increase (between the interdependent model and the basic model) is approximately 1.64-fold (a 43.078% increase). Similarly, the average increase in recall is approximately 1.70-fold (a 40.70% increase). When coupled as an F1-score, the average precision-recall increase is approximately 1.64-fold (a 38.55% increase). At some values of k ($K10$, $K50$ and $K100$) the interdependent model outperforms the matrix factorization model whereas at others ($K5$ and $K20$) the matrix model outperforms the interdependent model. On average however, the interdependent model still performs best (a 4.36% increase in precision compared to the matrix model, a 13.46% increase in recall and a 5.28% increase in F1-scores). It is also noted that the improvement of our interdependent model on the GC Twitter data is not as substantial as it is on the other two data sets. This is most likely due to the fact that the GC Twitter data set is not based on explicit check-in data, but rather, is implicitly derived under the assumption that a user is visiting a POI when they make a tweet nearby.



(a) Pre@K GC Twitter (b) Rec@K GC Twitter (c) F1@K GC Twitter

Fig. 6. Precision, recall and F1-scores on the GC Twitter data set.

Therefore, when considering model performance in terms of precision and recall accuracy, the interdependent model is considered very successful. In every data set analyzed, there is a significant increase observed. Whilst it is not as substantial with the GC Twitter data (when compared to the other data sets), it is still considered significant as a 5.28% increase in F1-scores. Such an excellent performance improvement is due to considering the coupling relationship between locations, and the inclusion of geographic information.

6 Conclusion

There are many limitations of traditional collaborative filtering models, mainly due to their poor performance on scarce POI data and lack of considering rich context information. A basic item-based collaborative filtering model is used to incorporate the interdependent location similarity to determine whether or not including this relationship improves prediction outcomes. In doing so, this study proposes a interdependent item-based model which successfully addresses the data scarcity and rich context issues by capturing the implicit relationship between the locations in terms of intra-similarity and inter-similarity. The interdependent item-based model also considers geographical data. The interdependent model displays excellent performance when used to analyze social network data sets. We observe excellent performance increases (up to 32-fold precision/recall increases and an 89% RMSE decrease) on a world-wide POI data set. Excellent performance is also observed when focusing on Australia and the Gold Coast social network data. The significant improvement verifies the great effectiveness of including the coupling relationship between locations, as well as the geographic information into the POI recommendation model.

The future directions of this research include optimizing the value of α with different data sets, considering what data characteristics may influence the coupled relationship between locations, and incorporating the coupled relationship into the advanced approaches for POI recommendation, e.g. deep learning model.

7 Acknowledgements

This work was supported by a Summer Scholarship project as well as two Griffith University's 2018 New Researcher Grants, with Dr Can Wang and Dr Md Saiful Islam being Chief Investigators, respectively.

References

1. Liu, Y., Pham, T.A.N., Cong, G., Yuan, Q.: An experimental evaluation of point-of-interest recommendation in location-based social networks. Proc. VLDB Endow. **10**(10) (2017) 1010–1021
2. Lu, Z., Dou, Z., Lian, J., Xie, X., Yang, Q.: Content-based collaborative filtering for news topic recommendation. In: AAAI 2015. (2015) 217–223
3. Ghazarian, S., Nematbakhsh, M.A.: Enhancing memory-based collaborative filtering for group recommender systems. Expert Systems with Applications **42**(7) (2015) 3801–3812
4. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8) (2009)
5. Yin, H., Wang, W., Wang, H., Chen, L., Zhou, X.: Spatial-aware hierarchical collaborative deep learning for poi recommendation. IEEE TKDE **29**(11) (2017) 2537–2551
6. Yin, H., Zhou, X., Cui, B., Wang, H., Zheng, K., Nguyen, Q.V.H.: Adapting to user interest drift for poi recommendation. IEEE TKDE **28**(10) (2016) 2566–2581
7. Liu, N.N., Yang, Q.: Eigenrank: a ranking-oriented approach to collaborative filtering. In: ACM SIGIR 2008. (2008) 83–90
8. Zhong, N., Li, Y., Wu, S.T.: Effective pattern discovery for text mining. IEEE TKDE **24**(1) (2012) 30–44
9. Thomas, A., Sindhu, L.: A survey on content based semantic relations in tweets. Int. J. Comput. Appl **132**(11) (2015) 14–18
10. Sattari, M., Manguoglu, M., Toroslu, I.H., Symeonidis, P., Senkul, P., Manolopoulos, Y.: Geo-activity recommendations by using improved feature combination. In: Ubicomp 2012. (2012) 996–1003
11. Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., Hinton, G.: Grammar as a foreign language. In: NIPS2015. (2015) 2773–2781